A CASE STUDY ON DETERMINING THE BIG DATA VERACITY: A METHOD TO COMPUTE THE RELEVANCE OF TWITTER DATA

By

JYOTSNA PARYANI

Bachelor of Engineering in Computer Engineering

Pune University

Pune, India

2012

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2017

A CASE STUDY ON DETERMINING THE BIG DATA

VERACITY: A METHOD TO COMPUTE THE

RELEVANCE OF TWITTER DATA

Thesis Approved:

Dr. K. M. George

Thesis Adviser

Dr. N. Park

Dr. Johnson Thomas

ACKNOWLEDGEMENTS

The Master degree from Computer Science department, Oklahoma State University has given me immense experience and knowledge in my field of interest. I would take interest to thank my Thesis Advisor and head of Computer Science Department, Oklahoma State University, Dr. K. M. George for his continuous assistance and encouragement to learn new technologies.

I would express my gratitude to committee members, Dr. N. Park and Dr. Johnson Thomas for their guidance and support. I would extend my thanks to my senior Ashwin Kumar Thandapani Kumarsamy for helping us in completing the thesis research.

Finally, I would express my profound gratitude to my family and friends who supported me throughout my years of study.

Name: JYOTSNA PARYANI

Date of Degree: MAY 2017

Title of Study: A CASE STUDY ON DETERMINING THE BIG DATA VERACITY: A METHOD TO COMPUTE THE RELEVANCE OF TWITTER DATA

Major Field: COMPUTER SCIENCE

Abstract: Twitter data (tweets) has all the attributes of Big Data. Also, it has become the source of information where people post their real-time experiences and their opinions on various day-to-day issues. Therefore, twitter data mining is being used for knowledge extraction and prediction in various domains. As its popularity and size grow, the veracity of knowledge extracted becomes a concern. Veracity is one of the V's of Big Data. The integrity of data, data authenticity, trusted origin, trustworthiness are some of the aspects that deal with Veracity. This thesis deals with the Veracity aspect of Big Data, in particular, veracity in Twitter data, from the truthful vantage point. In this research, we have compared existing Big Data Veracity models with a newly proposed measure. The proposed Veracity measure is entropy and it is compared with two other models, namely Objectivity, Truthfulness and Credibility model(OTC) and Diffusion, Geographic and Spam indices (DGS model) of Veracity. Our approach is to define topics on the set of tweets related to a domain and compute the veracity measures of the topics. The proposed model is based on the bag-of-words model for topic definition. Based on the values of the measures further inferences are achieved.

For our analysis, we selected three domains. The domains we chose are the flu, food poisoning, and politics. The topics for flu and food poisoning data are based on anchor words taken from CDC website. Anchor words of topics for Politics data are taken from "ontheissues.org" website. The entropy, OTC model, and DGS model are calculated for each topic. Our analysis shows no correlation between entropy, OTC model, and DGS model when compared as time series. Computed values of the models could position the topics in a veracity spectrum.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Micro-blogging sites like Twitter can be viewed as a social network or information network [1] [2]. Nowadays, Twitter is the most frequent platform for Big Data analysis [3]. It has become the source of information where people post their real-time experiences and their opinions on various day-to-day issues which can be used to predict and analyze the data. In an International Data Corporation report, it was predicted that "from now until 2020 the data will double in every 2 years", therefore resulting into Exabyte of data [4]. This leads to a challenge of security and trust of the data available on Social networking sites.

The integrity of data, data authenticity, trusted origin, trustworthiness are some of the aspects that deal with Veracity of data [5]. According to IBM Big Data & Analytics Hub [6], 27% of the respondents were not sure about the accuracy of the data and one in three decision-makers do not trust the information used for analyzing the data. Along with the above aspects, there are some characteristics of Big Data like Volume, Variety, and Dynamicity which constitutes for Big Data security issues [5]. With the increase in the size of data and the myriad variety of data, the data available for analysis should be trustworthy, not outdated or manipulated. The dynamicity [5] deals with the change in structure, data model and migration of datacenter, which brings into the picture the confidentiality and integrity of data. Moreover, the value of the data can be hidden in jargon or linguistic which may result in the information not recorded in writing or may mislead the recipient [7]. The velocity of the data can also deal with removing or altering the important information which may, in turn, affect the trustworthiness of data.

Therefore, Veracity of Big data depends on many other characteristics of Big Data like Volume, Variety, Value and Velocity.

Table 1 lists some V's of Big Data and their definitions [8] [9]

Table 1: Definition of V's in Big Data

Sr. No	Characteristics of Big Data	Description of Characteristic
1	Volume	Scale of data [6].
2	Variety	Different forms of data. (unstructured data constitutes of 80% of world data) [6]
3	Velocity	Analysis of Streaming data [6].
4	Value	Extracting business value from the data [9].
5	Veracity	Uncertainty of data [6].

This thesis deals with the Veracity aspect of Big Data from the truthful vantage point. We are interested in estimating veracity from the data without relying on external information as it may not be feasible to collect and analyze external information. There is some research conducted in this area such as Objectivity, Truthfulness, and Credibility (OTC) model [10] and measures of Veracity [11]. In this thesis, we propose entropy as another measure of Veracity and develop a method to associate entropy to topics defined as a bag-of-words. The entropy measure is compared against the previously defined measures. There are two motivations for our approach. First, no measure available is 100% accurate in determining the truthfulness of data and so more measures are needed for validation. Second, entropy is used to measure the ambiguity in statements which may be interpreted as a measure of truthfulness in tweets. We apply the measures to topics built on Twitter data.

The analysis is done on different kinds of datasets like the flu, food Poisoning, and Politics. These datasets were chosen for the following reasons.

Starting with flu dataset, there have been many flu pandemics throughout the history. The study of CDC has concluded that each year 200,000 people in the United States are hospitalized each year for respiratory and heart conditions illness associated with seasonal influenza virus infections (refer to Table 19, row 3). Moreover, CDC has estimated that from the 1976-1977 season to 2006-2007 flu season, flu-associated deaths range from as low as about 3,000 to as high as about 49,000 people (refer to Table 33, row 3). Therefore, the correctness of information related to flu in social media is a vital factor for society. Flu data is collected from Twitter using keywords collected from CDC (refer to Table 19, row 3 and row 4), Mayo Clinic (refer to Table 19, row 9), Flu.gov (refer to Table 19, row 6 and row 7) and WebMD (refer to Table 19, row 18, row 19 and row 20) links. Further, the data is classified based on flu topics and then statistics and evaluation are done based on the flu data.

Next, the analysis is done on food poisoning data as the study of CDC has estimated that every year roughly 48 million people get sick from a foodborne illness, 12800 are hospitalized and 3,000 die from foodborne illness (refer to Table 19, row 23). Therefore, the precision of information related to food borne illness in social media is a vital factor for society. Food Poisoning data is collected from Twitter using keywords listed in CDC website (refer to Table 19, row 24). Similarly, like flu data, the data is classified based on food poisoning topics and then statistics and evaluation is done based on the food poisoning data.

Lastly, the analysis is done on Politics data based on US elections 2016 results related to US president Donald Trump. The data is collected from Twitter using common keywords like Donald Trump, immigration, Muslim, terrorism, Mexico. The analysis is done on politics topics and statistics and evaluation is done based on politics data

CHAPTER II

LITERATURE REVIEW

2.1 RELATED WORK

The vastness and diversity in data have opened the new opportunities in Big Data but on the same hand, it has lead into questions in trust of various parts dealing with Big Data like collection and preparation of data, storage of data, data quality, nodes, Cloud Service Providers and information sharing [12]. Big Data is initially characterized by the 3V's Volume, Variety, and Velocity. As it evolves, more V's are added to this characterization; the other V's are Veracity, Value, etc. Veracity deals with the reliability of data. Several factors contribute to the reliability of data. This thesis deals with the trust factor of the data. The research is based on different works proposed in the literature. In this chapter, we review the previously published research related to and contributing to our research. These works can be classified as topic modeling, sentiment computations, information theory, and veracity. Research related to each of the above works are summarized in the following subsections.

2.1.1 TOPIC MODELLING

Topic modeling refers to a generative model for analyzing large quantities of unlabeled data. A topic is a probability distribution over the collection of words and topic model is the statistical relationship between a group of observed and unknown random variables that specifies a probabilistic procedure to generate the topics [13]. One of the most popular topic modeling technique used is Latent Dirichlet Allocation (LDA) [13].

Latent Dirichlet Allocation (LDA) is a generative probabilistic model which extracts the topic in the text, based on co-occurrence of words with the topic in the document [14] [13]. Following is the algorithm of LDA [14] [13]:

For each document w in a corpus D:

1. Choose N ~ Poisson (ξ).
2. Choose Θ ~ Dir (α), where Dir (.) is a draw from a uniform Dirichlet distribution with scaling parameterα.
3. For each of the N words wn:
    a. Choose a topic zn ~ Multinomial (Θ).
    b. Choose a word wn from p (wn | zn,β), a multinomial probability conditioned on the topic zn .

There are many other similar models and techniques related to topic models like Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Non-negative matrix factorization (NMF) and Correlated Topic Model (CTM). LSA is a statistical technique which deals with extracting and representing the relations between words in a large corpus. This technique of LSA helps in information retrieval from a large text [15]. PLSA which evolved from LSA is a probabilistic generative model which associates unobserved variables with each occurrence of a word in a document [16]. This co-occurrence of words in the document has applications in information retrieval and filtering, machine learning from text and natural language processing [17]. Next model, NMF a document clustering method, deals with finding the latent semantic structure for the document corpus and identify document clusters in the derived latent semantic space. In the latent semantic space, each axis represents the basic topic of a document cluster and each document is represented as a combination of the base topics [18]. One of the limitations in LDA topic modeling technique is addressed in Correlated Topic Model. LDA is unable to model topic correlation between the generated topics from the model. CTM generates the topic graph where each node indicates the topic (containing the set of most probable keywords), the font of the node represents its popularity and lastly correlation with other nodes [19].

2.1.2 INFORMATION MEASURE

In this thesis, we adopt entropy as a measure for veracity. It is proposed by Claude Shannon known as Shannon's mathematical theory of communication [20]. Shannon's theory deals with information to be conveyed with three communication problems such as first, the accuracy of the information to be transmitted; second, how precisely the meaning is transferred and third, from all the information transferred how much is selected from the set of messages. The last aspect is basically the effectiveness of the information transmitted from the sender to the receiver. The information in this context basically deals with a message to be selected from the set of messages. From the set of messages, there should be a function to choose a message from the set. This selection process can be done with the help of logarithmic function (base 2) [21]. The logarithmic function is used because if the set of messages increase from 2 choices to 8 choices, the logarithmic measure increases from 1 bit to 3 bits of information [20]. Moreover, from the set of possible events or messages, the probability of occurrence of each event is attached with the logarithmic function. Finally, the individual probability and logarithm of an event are then added to get the information measure as Entropy from the set of events or messages.

Shannon's Entropy [20] is, therefore, the information required to describe an event or entity. Following is the Entropy equation:

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

Where the n= number of different outcomes.

The range of Entropy is $0 \leq \text{Entropy} \leq \log(n)$ which is based on a number of outcomes. Maximum entropy ($\log n$) occurs when all the probabilities have equal values that are 1/n. Minimum entropy (0) occurs when one of the probabilities is 1 and rest are 0's [20].

2.1.3 BIG DATA VERACITY: SENTIMENT LIBRARY

There are very few Veracity models available in the literature, one of which is the Big Data Veracity model: Objective, Truthful and Credible (OTC) [10]. It will be used to compute the Veracity measure and compared against the entropy model that we propose in this thesis. Details of the OTC model are given below.

2.1.3.1 Objectivity

OTC is a three-dimensional model proposed by Lukoianova. The first dimension of Veracity in OTC model is objectivity/subjectivity of the data. Objectivity basically deals with facts [22], truth, reality and reliability [23], [24]. Objectivity is the knowledge which is proven whereas subjectivity is the knowledge which is weakly supported or has the error possibility [23] [24]. There are linguistic tools like automatic essay scoring, automatic classification of the sentiment or opinion expressed in a text, which scores the sentiment or opinion classification expressed in the text [25]. Moreover, objectivity is just a version of the truth and therefore veracity also deals with next dimension that is truthfulness/deception [22] [10].

2.1.3.2 Truthfulness

The second dimension, truthfulness in the textual data can be determined by checking if there is any false belief or false conclusion in the text [26] [27], which can be verified with the help of deception test. If the test is passed, then it is the truthful text but if the test is failed then the user has to further look into alternatives and dig deeper for further fact verification [10]. Also, with the increasing volume of data, there are more chances of deceptive text communication [10]. Therefore, more the text is deceptive, it would lead to incorrect analysis and results. There are many deception detection software which calculates the statistics used for classifying truthful or deceptive text (refer to Table 19, row 11) or identify fake product/service reviews on the websites (refer to Table 19, row 11) [28] [29]. The advantages and disadvantages of the above-mentioned deception tools are evaluated in [30].

2.1.3.3 Credibility

The last dimension of the OTC model is the credibility of the data. Credibility deals with two qualities, trustworthiness and expertise. Trustworthiness deals with qualities such as truthful, well-intentioned or unbiased. Expertise deals with knowledgeable, reputable and competent qualities in the related field [31]. Trustworthiness of the content can be achieved by relying on character, ability, strength or truth of trusted content (refer to Table 19, row 10). The above mentioned two qualities of credibility can be achieved either through the vocabulary of trust or credible source [10]. The credibility of the text can be calculated using Mutual Information between words which consists of performing analysis on frequently occurring nouns and verbs with the trust and credibility [10] using online corpus COCA [32].

2.1.3.4 Calculating OTC model using Sentiment library

We have calculated Objectivity and Truthfulness using text processing library 'Text Blob' in Python (refer to Table 19, row 2). Text Blob (For installation & downloading (refer to Table 19, row 5 and row 8)) library is used to perform natural language processing tasks such as parts-of-speech tagging, noun phrase extraction, sentiment analysis, classification, language translation, word tokenization and many more features (refer to Table 19, row 14). The sentiment property of Text Blob calculates the subjectivity and polarity of the given text (refer to Table 19, row 2). Subjectivity is the negative side of Objectivity in the OTC model. The calculated subjectivity is a floating point number in the range of [0.0, 1.0] where 0.0 means the text is very objective and 1.0 means the text is very subjective (refer to Table 19, row 2). The polarity deals with positive, negative and neutral observed emotions in the text. These emotions can be used to differentiate between truthful and deceptive text [33] [34] [35]. The evidence that the liars use more negative emotions than the truth-tellers [36] can be used as a measure to find the polarity in the text, which is the second dimension that is Truthfulness/Deception in OTC model. The polarity calculated in Text Blob library is a floating point number in the range of [-1.0, 1.0] where 1.0 represents very negative emotion, 1.0 represents very positive emotion and 0.0 represents neutral emotion (refer to Table 19, row 15).

The Text Blob library depends on NLTK and Pattern libraries (refer to Table 19, row 16). NLTK library has corpus like WordNet which is used by the library and by default Text Blob uses Pattern Analyzer (refer to Table 19, row 2). Pattern Analyzer is a sentiment analyzer that uses Pattern Library (refer to Table 19, row 17). Pattern is a library in Python which performs data mining operations like natural language processing, machine learning, network analysis, and visualization. This pattern library is created by 'Tom De Smelt and Walter Daelemans' [37] (refer to Table 19, row 12) and Tom De Smelt has performed various case studies on python where he has given the detailed approach for Sentiment Analysis ([38] Chpt 7, p. 133). The sentiment analysis consists of Lexicon of Dutch adjectives, used by Pattern Library, which consists of the list of adjectives and their polarity, subjectivity and intensity scores. This Dutch lexicon is then converted into English lexicon ([38], Chpt 7) and each adjective is associated with properties such as wordnet_id, cornetto_synset_id, pos(parts of speech), sense(meaning of the word), polarity, subjectivity and intensity scores (refer to Table , row 13).

The Dutch Lexicon is created in three steps as manually annotating 1000 adjectives, performing one semi-supervised and another supervised machine learning method. Initial adjectives are taken from online Dutch book reviews and 7 human annotators have given polarity and subjectivity scores to the adjectives. Then scores are compared with 'DUOMAN subjectivity Lexicon for Dutch' [39], which uses Page Rank Algorithm [40] to annotate the adjectives. The second step is to expand the list of adjectives, which is done by taking the adjectives from Dutch Newspaper Corpus. The approach taken is similar to using a vector space with adjectives as labels and nouns as feature vectors [41]. Then, for each adjective which is both in newspaper corpus and initial lexicon, the number of times the adjective precedes the nouns is counted which results is adjective vectors for each noun and nouns as feature vectors. Then for each adjective in the initial lexicon, K-NN using cosine similarity [42] is applied to retrieve 20 most similar nearest neighbors from the newspaper corpus. The scores of initial lexicon are inherited by the newly discovered adjectives which result in 3200 adjectives ([38] Chpt 7, p. 135). The third step is to find the synsets from CORNETTO (Dutch WordNet) for each adjective found in the previous step. The synsets are the relations (synonym, antonym) between the words in CORNETTO. After retrieving new adjectives from CORNETTO, the scores of adjectives are again passed on to new adjectives resulting in 5400 adjectives ([38] Chpt 7, p. 136).

Lastly, the Dutch adjective lexicon is converted into English lexicon using inter-language relations in CORNETTO, having reference to WordNet. More adjectives were taken from IMDB movie reviews and then added to English lexicon by a manual single annotator ([38] Chpt 7, p. 137) and then compared with Polarity Dataset 2.0 [43].

After computing Objectivity and Truthfulness, Credibility is computed using mutual information between 2 words. The mutual information deals with co-occurrence between 2 words. First, the words are tokenized using NLTK word tokenize function, then the number of words (n) is calculated in the text and the sample space taken is n*n. Next, the count of a number of times each of the 2 words are present in the text is computed. Let these count be n1 & n2.

$$\text{Mutual Information} = \frac{\text{probability of word1 \& word2}}{\text{probability of word1} * \text{probability of word2}}$$

$$\text{Therefore, Mutual Information} = \frac{\mathrm{n1} * \frac{2}{\mathrm{n}^2} + \mathrm{n2} * \frac{2}{\mathrm{n}^2}}{\frac{2\mathrm{n}-1}{\mathrm{n}^2}}$$

Finally, all three dimensions of the OTC model are combined into one Veracity index, by normalizing the dimensions in the range of (0, 1) interval with 1 being maximum objectivity, truthfulness, credibility and 0 being minimum objectivity, truthfulness, credibility [10]. This composite index provides a way of assessing systematic variations in big data quality across datasets with textual information. The index could be helpful to identify those parts of the big dataset that are of lower quality for their subsequent exclusion if the quality of the entire dataset is to be improved [10].

2.1.4 BIG DATA VERACITY: QUANTATIVE MEAURES

Another big data veracity model in the literature is proposed in [11] which deals with evaluating the accuracy of data from the tweets. The objective of this model is to compute veracity measures as external resources may not be readily available for verification from the data. The model proposes three measures based on the spread of information in term of volume, geographic spread and repetition in the volume. It is based on the argument that information with high volume and inflation rate spreads widely and could be questionable. All the measures are defined with tweets as the source of data.

The three measures proposed are Topic Diffusion, Geographic Dispersion and Spam Index [11]. The first measure, Diffusion Index deals with how fast the information has spread through Twitter. It deals with the concept that fast information has spread faster than the truth. [44].

$$\text{Diffusion Index} = \frac{\#\text{ Unique Users}}{\text{Total tweets}}$$

The second measure, Geographic Spread Index is used to measure the extent to which the information is spread geographically.

$$\text{Geographic Spread Index} = \frac{\#\text{ Unique Location}}{\text{Total tweets}}$$

And the last measure is the Spam Index, which deals with the impact of repeated tweets by the same user. Repeated tweets can be viewed as inflating the diffusion. The measure is similar to spamming which propagates questionable information [11].

$$\text{Spam Index} = \frac{\sum_{\text{over unique users}} \frac{1}{\text{unique user tweet count}}}{\text{Total tweets}}$$

For convenience, we call this the DGS model. This model is based on the set of tweets and selected content of tweets.

2.1.5 OTHER WAYS OF COMPUTING BIG DATA VERACITY

There are various other ways of improving data veracity like performing masking on the data prior to inserting the data in database and ensuring that only authorized users can unmask the data [45]; in smart electric grids having geographically dispersed sensors for real-time modelling and decision making can be done from only small subset of influential sensors and making predictions for all sensors [46]; Veracity(trustworthiness) of web event is measured based on uncertainty of attributes, uncertainty of webpage and website confidence [47]. One of the ways of improving Veracity is described in “Learning from Uncertainty for Big Data” [48] which deals with various definitions of uncertainty like Shannon entropy, Classification entropy, Fuzziness and Nonspecificity (Ambiguity).

There are papers [49] relating to junk data in data collection that are not reviewed for this research. Our focus is on the validity and truthfulness than the error in data collection.

2.2 PROBLEM STATEMENT

As outlined in previous sections, Veracity is an important aspect of Big Data and especially twitter data. This thesis proposes entropy as a measure of Veracity and compare its reliability against the previously published measures. Our approach to analyzing veracity of Twitter data is to define topics on the tweets and apply veracity measures to the topics. We conjecture that determining the veracity in absolute terms will be difficult to accomplish based on tweets only. So, we propose that veracity be considered as a spectrum. Models are viewed as locators of topics within the spectrum.

CHAPTER III

METHODOLOGY

3.1 TOOLS USED

This section includes the tools used for data collection

3.1.1 APACHE HADOOP

Apache Hadoop (Table 19, row 1) is an open-source software for reliable, scalable and distributed computing. The software library is a framework that allows for distributed processing of large volume of data across clusters of computers using simple programming models. Hadoop is designed to expand from single machines to thousands of machines and each machine offering local computation and storage. The Hadoop project includes Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN and Hadoop MapReduce. Hadoop Common are the set of utilities which support other Hadoop modules. HDFS (Table 19, row 21) is the distributed file system which is fault tolerant and is deployed on low-cost hardware. It is suitable for applications having large data sets and provides high-throughput access to data. Some of the goals of HDFs are fault detection, quick and automatic recovery, batch processing, coherent and portable across heterogeneous platforms.

3.1.2 APACHE FLUME

Apache Flume (Table 19, row 22) is a robust and fault tolerant distributed service used for collecting, aggregating and moving large amounts of data. It has a simple and flexible architecture based on data streaming flows (refer to Figure 1).

Following is the Twitter Data Streaming Process: To start the streaming, Flume uses its integral components such as agent, source, channel, sink, and event.

1. Source connects to the source of data (Twitter) and sends the data to the sink through the channel.
2. Channel acts as a bridge between Source and Sink.
3. The data from the final stage that is sink is transferred to HDFS.
4. An event is the basic unit of data that is transferred using flume.
5. An agent is a container for data flow.

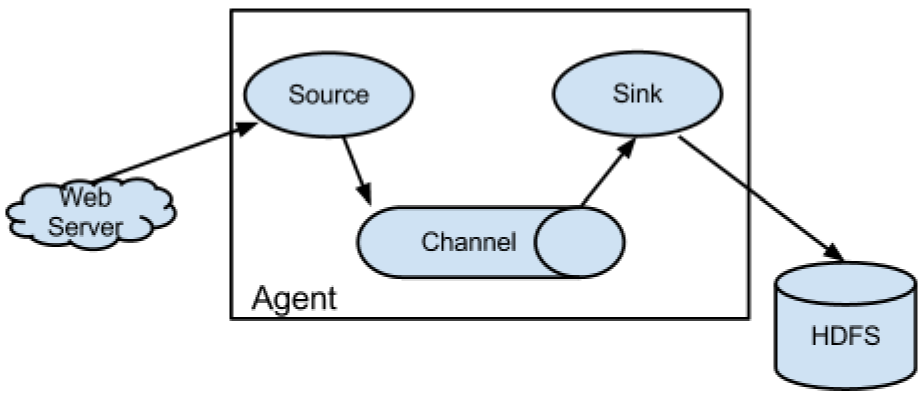


Figure 1: Flume Agent

3.2 DATA COLLECTION

Three different types of data are collected. First, we have collected 216 GB of Flu Data. The data collection period is 06/1/2015-11/30/2015. The tweets are collected using set of keywords like Fever, Feverish chills, chills, Cough, Sore Throat, Runny Nose, Stuffy Nose, Body Ache, Muscle Ache, Headache, Fatigue, Tiredness, Tired, Vomiting, Diarrhea, Joint Aches, pain around eyes, watery eyes, flushed skin, exhaustion, sneezing, dizziness, runny nose, stuffy nose, cough, diarrhea, headache.

Second, we have collected 18 GB of Food Poisoning Data. The data collection period is 11/1/2016-01/15/2017. The tweets are collected using set of keywords like diarrhea, fever, abdominal pain, abdominal cramps, vomiting, bloody diarrhea, muscle ache, nausea, headache, stiff neck, confusion, convulsion, chills, watery diarrhea, stomach cramps, weight loss, slight fever, greasy stools, gas and bloating, double vision, blurred vision, drooping eyelids, slurred speech, difficulty swallowing, dry mouth, muscle weakness, jaundice, dark-colored urine, light-color stool

Third, we have collected 200 GB of Politics data. The data collection period is 2/17/2017 – 3/17/2017. The tweets are collected using the set of keywords like trump, Donald, immigration, Muslim, terrorism, Mexico.

Apache Flume is used to retrieve data from tweet stream. For streaming the data, we have created flume agent and twitter application. The twitter application contains the set of keywords related to the flu. From the application, API keys are used for streaming the data in Hadoop cluster. For flume agent, the configuration file is created which contains tokens of the twitter application. The data obtained from twitter is in JSON format.

3.3 DATA PREPROCESSING

From the JSON format file, tweet text, created date, geolocation and user fields are retrieved for further processing. The tweets text is then cleaned by removing the URL’s, user mentions, internet slang words, emoticons and stop words.

3.4 EVALUATION MEASURE

The Veracity of Big Data deals with the uncertainty of data. In this research, we use Twitter data as a case study tool. This research proposes Entropy as a measure to evaluate the veracity topics using contributing tweets. Shannon’s Entropy can be used to measure the ambiguity in the information contained in a text. We interpret this ambiguity as for the measure of veracity implicit in the tweets.

The Entropy is calculated based on Shannon’s Entropy formula:

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

The results of Entropy measure are then compared with two other models from the literature. The models used for comparison are the OTC Model and Diffusion, Geographic and Spam Index (DGS) model.

CHAPTER IV

FINDINGS

4.1 TOPIC EXTRACTION

This section deals with presenting the result of analysis from Shannon's Entropy, OTC Model, and DGS model by applying to several topics. A topic is defined by a set of keywords (or anchor words) and a document may consist of several topics. Topic Extraction deals with extracting information from documents. Topic modeling is a popular method to identify topics. Topic modeling refers to a generative model for analyzing large quantities of unlabeled data. LDA is a popular technique of topic modeling. LDA is a generative probabilistic model which groups similar keywords under a topic based on co-occurrence of words with the topic in the document.

Another way of Topic Extraction can be getting the topics by manual extraction of keywords belonging to them. In this research, we have chosen the manual method of extracting topics from the CDC Website for Flu and Food Poisoning data. The topics for Politics data are extracted from the "on the issues" website for Donald Trump. We found the topic modeling approach inefficient and inaccurate to our research. The Big Data Veracity measures are then evaluated for a topic based on tweets containing the set of keywords in a topic. One reason for us to adopt the topic extraction approach is the computationally expensive nature of LDA as shown in the next section.

4.1.1 LDA Topics & Performance

In this section, we analyze the performance of the LDA algorithm. The LDA algorithm is implemented by using the parameters document term matrix of the text, a number of topics and Gibbs sampling for computing posterior distribution of words assigned to a particular topic. We have executed the LDA algorithm on different sizes of data and computed the CPU time on all the sizes. The software used for executing the LDA algorithm is R and the results were executed on CSX server. The data sizes are from 8 bytes to 3.5 GB of data. A multi-line graph is plotted to depict the LDA performance by taking logarithm of data size versus CPU time. Table 2 describes the data sizes and corresponding CPU times for running the algorithm. The multi-line graphs in Figure 2 compare user and system CPU times. User CPU time deals with actions performed on program and system CPU time deals with the time spent in performing system calls for the kernel.

Table 2: Data Size & CPU Time

Data Size	CPU Time (user, system) in secs	Data Size	CPU Time (user, system) in secs
23 = 8 Bytes	0.027,0.000	218 = 256 KB	6.275, 0.002
24 = 16 Bytes	0.027,0.000	219 = 512 KB	12.704, 0.005
25 = 32 Bytes	0.029,0.002	220 = 1 MB	25.125, 0.009
26 = 64 Bytes	0.031,0.001	221 = 2 MB	50.672, 0.046
27 = 128 Bytes	0.033,0.000	222 = 4 MB	103.812, 0.076
28 = 256 Bytes	0.037,0.001	223 = 8 MB	223.625, 0.224
29 = 512 Bytes	0.047,0.001	224 = 16MB	271.766, 0.324
210 = 1 KB	0.064,0.002	225 = 32MB	554.492, 0.413
211 = 2 KB	0.102,0.001	226 = 64MB	1638.428, 1.52
212 = 4 KB	0.081,0.000	227 = 128MB	3028.683, 3.963
213 = 8 KB	0.265,0.000	228 = 256MB	24142.294, 25.555
214 = 16 KB	0.488, 0.000	229 = 512MB	27925.940, 75.511
215 = 32 KB	0.971, 0.000	230 = 1GB	68304.798, 367.11
216 = 64 KB	1.737, 0.000	231 = 2GB	272082.48, 1134.22
217 = 128 KB	3.363, 0.001	232 = 4GB = 3.5GB	1009656.762, 10422.903

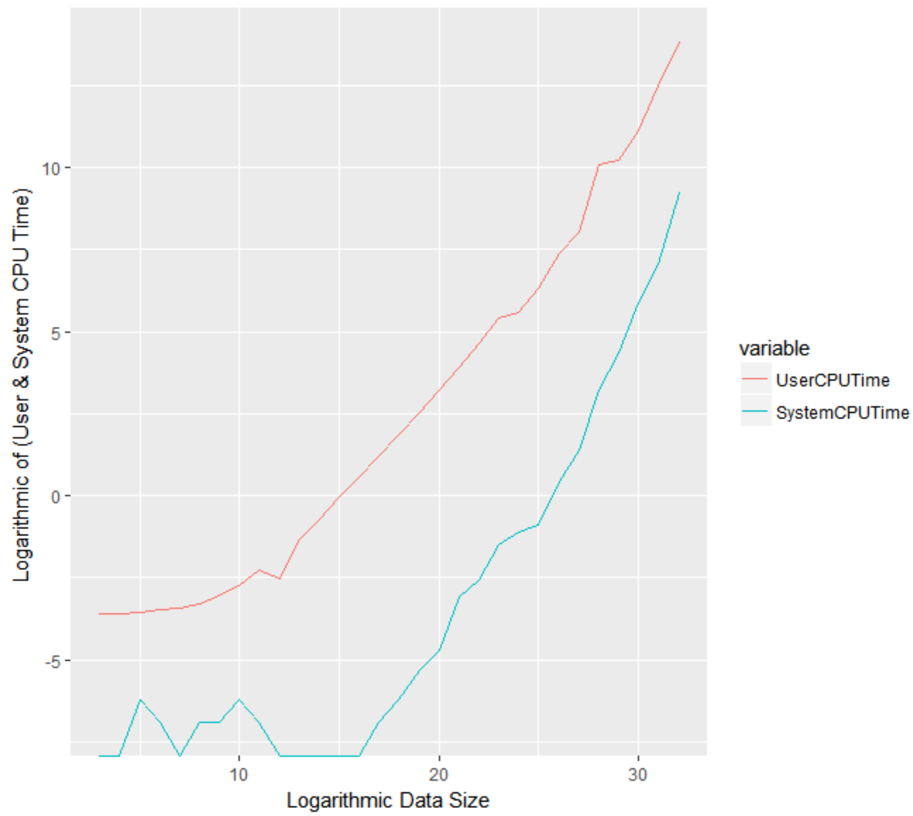


Fig 2: Line Plot of Data v/s CPU time

Based on the results, the following conclusion can be derived:

- User CPU Time is much higher than System CPU Time.
- Minimum Value for User CPU Time is much more than System CPU Time
- The System CPU Time is constant for values between 12 to 16 log (Data Size) and then increases as Data Size increases.

4.1.2 TOPIC EXTRACTION BY MANUAL METHOD

There are two motivations for using the topic extraction method, one is the expensive nature of topic modeling algorithm (LDA) which is described in the previous section and the second reason is to divide the set of keywords based on the different context of the data which is not done in LDA. For example, various contexts of flu data are symptoms, side-effects and treatments and the keywords related to the context are grouped in one topic. These topics are retrieved from the CDC website.

1. Topic 1: Fever, chills, Feverish chills, Sneezing, Body Ache, Muscle Ache, Weakness, Stuffy Nose, Diarrhea, Vomiting, Cough, Sore Throat, Flushed skin, Runny Nose, Nasal Congestion, Tired, Tiredness, fatigue, Headache (Flu symptoms)
2. Topic 2: Nausea, Vomiting, Delirium, Headache, Muscle Ache, Itching, Runny Nose, Nasal Congestion, Fever, Soreness, Redness, Swelling, cough, aches, fatigue, hoarseness, Diarrhea, sinusitis, Dizziness, bronchitis (Flu Side-effects)
3. Topic 3: Rest, medicine (treatments)

The second dataset used in our experiments is food poisoning. Various contexts of food poisoning data are symptoms of bacterial foodborne germs, symptoms of viral foodborne germs and symptoms of parasitic foodborne germs. The keywords related to various contexts are also taken from the CDC Website:

1. Topic 1: Symptoms of Bacterial Foodborne germs - double vision, blurred vision, drooping eyelids, slurred speech, difficulty swallowing, dry mouth, muscle weakness, Fever, chills, Headache, Nausea, Vomiting, Body aches, cough, dizziness, tiredness, sweats, Hoarseness, Fainting, Swelling of abdomen, flushing, Fainting, sore throat, malaise, anorexia, Fatigue, pain in muscles, joint, and/or back, depression, low blood pressure, thirst, muscle cramps, restlessness, rapid heart rate, loss of skin elasticity, dry mucous membranes, abdominal

cramps, diarrhea, weakness, anemia, Rash, Red eyes, Jaundice, loss of balance, stiff neck, confusion, Tenesmus.

2. Topic 2: Symptoms of Viral Foodborne germs – diarrhea, throwing up, Nausea, stomach pain, fever, headache, body aches, dry mouth and throat, feeling dizzy, sleepy or fussy, cry, Fatigue, Abdominal pain, Dark urine, Jaundice, vomiting, Loss of appetite, Clay-colored bowel movements, Clay-colored stool
3. Topic 3: Symptoms of Parasitic Foodborne germs - stomach pain, stomach cramping, bloody stools, fever, abdominal pain, nausea, vomiting, abdominal distention, diarrhea, blood and mucus in stool, abdominal discomfort, Weight loss, Dehydration, Stomach cramps or pain, Watery diarrhea, bloating, loss of appetite, Gas, Greasy stools, reduced vision, blurred vision, pain (often with bright light), redness of the eye, muscle pains, itchy skin, constipation, heart and breathing problems, swelling of the face and eyes, cough, chills.

The third dataset used in our experiments is data from the political domain. Various topics/concepts of interest in the politics domain are social, economic, domestic and international issues. We extracted keywords and phrases related to these topics from the website "ontheissues.org".

1. Topic 1: Social Issues – abortion, civil rights, education, families & children, welfare & poverty, principal & values.
2. Topic 2: Economic Issues – budget & economy, corporation, government reform, tax reform, social security, jobs.
3. Topic 3: Domestic Issues – crime, drugs, gun control, health care, technology, environment.
4. Topic 4: International Issues: foreign policy, homeland security, war & peace, free trade, immigration, trade & oil.

In the subsections that follow, we provide the results of the three veracity measures OTC model, Entropy model, and DGS model applied to the above-mentioned domains. In each domain, topic related data from among the tweets are retrieved. Based on this data, the three measures are computed for every topic in the domains. The topics are ranked based on the scores and then ranking of topics for entropy measure is compared with the ranking of topics for the OTC model and with the DGS model.

4.2 BIG DATA VERACITY: ENTROPY

In this section, we present the results of evaluation of the Entropy measure applied to the flu, food poisoning and politics data. As mentioned previously, the Entropy measure is defined using Shannon's Entropy formula:

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

where p_i represents the probability associated the i^{th} keyword defining the topic. In our computation, keyword probabilities are computed by the formula $p_i = \frac{n_i}{N}$ where N is the total number of words obtained from related tweets after excluding stop words and other insignificant words and n_i is the number of occurrences of the i^{th} keyword.

Tables 3 shows the computed Entropy scores for the flu topics (refer to section 4.1.2), Food poisoning topics (refer to section 4.1.2) and the Politics Data topics (refer to section 4.1.2) respectively. Figures 3, 4 and 5 show their respective histograms.

We proposed the Entropy Model as a measure of veracity. If the score is high, then there will be a higher degree of ambiguity which shows less certainty and less veracity. If the Entropy score is lower, it will show a high degree of certainty and thus higher degree of veracity. Based on this we can estimate the degree and ordering of the veracities of the topics.

Table 3: Entropy Score

Topics	**Flu Data Entropy Score**	**Food Poisoning Data Entropy Score**	**Politics Data Entropy Score**
Topic 1	0.4165	0.4041	0.3512
Topic 2	0.4126	0.4098	0.3266
Topic 3	0.3606	0.4089	0.3403
Topic 4	-----	------	0.3456

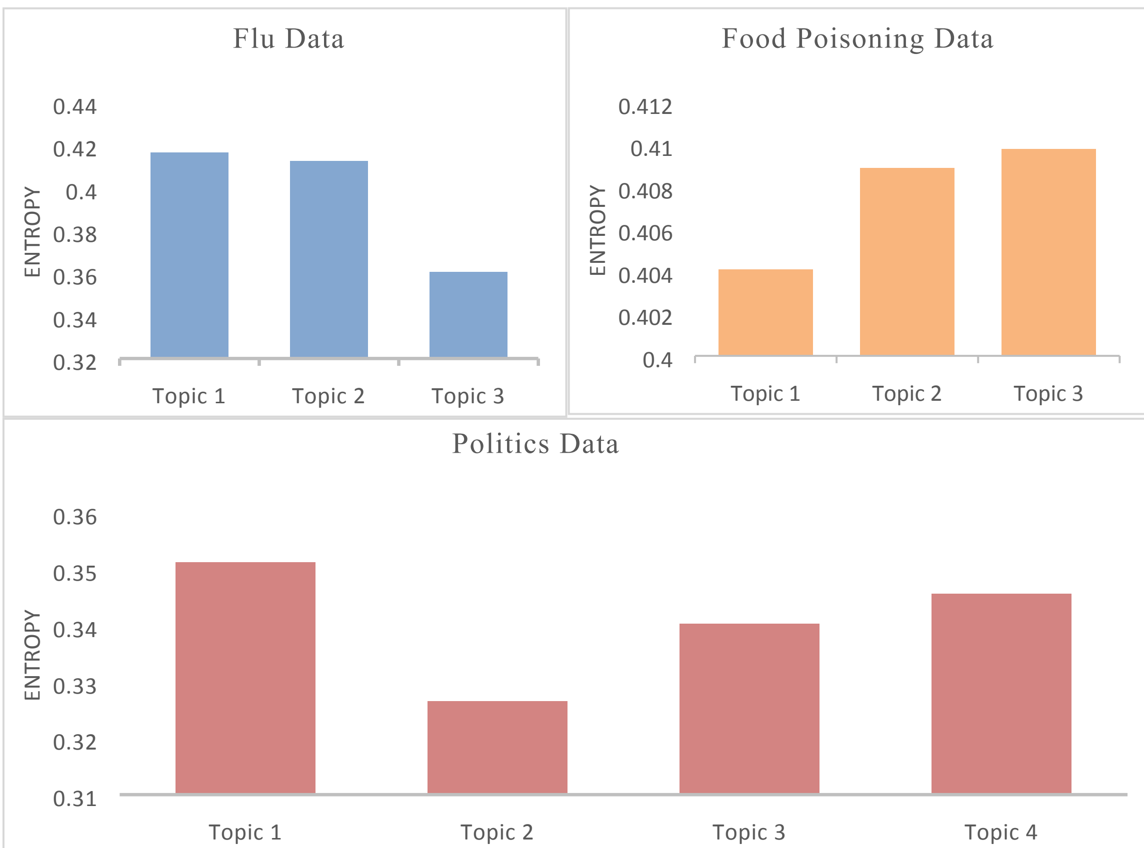


Figure 3: Histogram of flu data entropy score, Figure 4: Histogram of food poisoning data entropy score and Figure 5: Histogram of politics data entropy score

4.3 BIG DATA VERACITY: OTC MODEL

In this section, we present the results of evaluation of the OTC model applied to the flu, food poisoning and politics data. As mentioned previously, the objectivity and truthfulness measure of the OTC model is computed by the TextBlob library and credibility of the model is defined using mutual information between two words and it is defined by the following formula (refer to section 2.1.3):

$$\text{Mutual Information} = \frac{\text{probability of word1 \& word2}}{\text{probability of word1} * \text{probability of word2}}$$

Table 4 shows the individual objectivity, truthfulness and credibility scores for flu data. The range of objectivity and credibility score in the OTC model is [0,1]. The range of truthfulness score in the OTC model is [-1.0, 1.0]. After computing the truthfulness of the model, the score is normalized in the range of [0, 1] (refer to Table 5) and then the average of all the 3 scores is computed (refer to Table 6). If the OTC score is high, then there will be a higher degree of certainty and a higher degree of veracity. If the OTC score is lower, it will show a higher degree of uncertainty and thus lower degree of veracity. Based on this we can estimate the degree and ordering of the veracities of the topics.

Tables 6 show the computed OTC model scores for the flu topics (refer to section 4.1.2), Food poisoning topics (refer to section 4.1.2) and the Politics Data topics (refer to section 4.1.2) respectively. Figures 6, 7 and 8 show their respective histograms.

Table 4: Flu data OTC Score

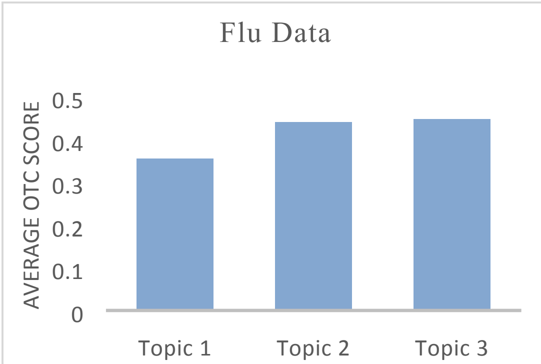
Topics	Objectivity Score	Truthfulness Score	Credibility Score
Topic 1	0.5317	-0.1747	0.1198
Topic 2	0.7170	-0.0027	0.1072
Topic 3	0.6405	0.1242	0.1328

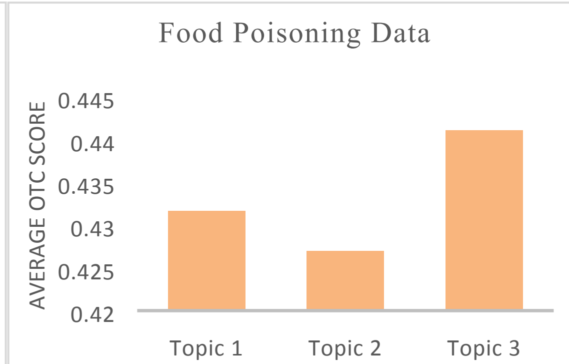
Table 5: Flu Data Normalized Truthfulness Score

Topics	**Flu Data Normalized Truthfulness Score**
Topic 1	0.4126
Topic 2	0.4986
Topic 3	0.5621

Table 6: Average OTC Score

Topics	**Flu Data OTC Score**	**Food Poisoning Data OTC Score**	**Politics Data OTC Score**
Topic 1	0.3547	0.4317	0.4429
Topic 2	0.4407	0.4270	0.4480
Topic 3	0.4473	0.4411	0.4332
Topic 4	-----	------	0.4443





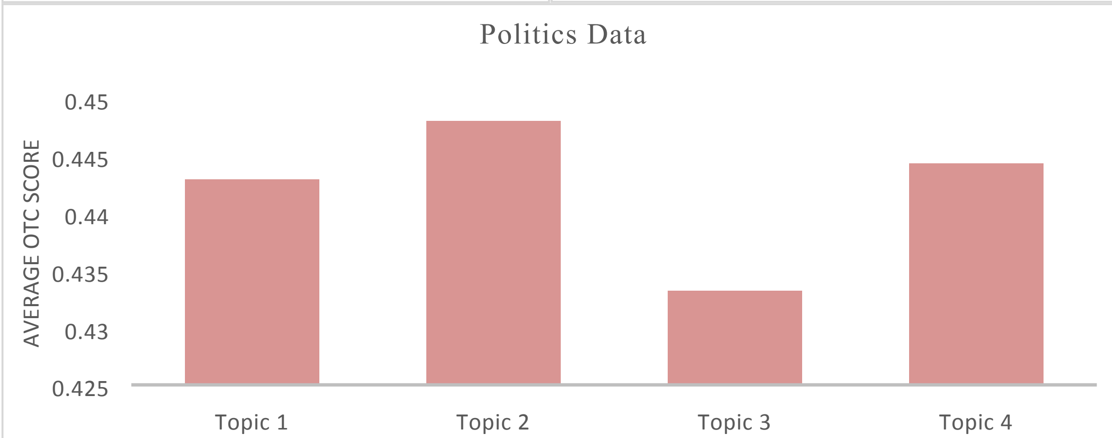


Figure 6: Histogram of flu data OTC score, Figure 7: Histogram of food poisoning data OTC score and Figure 8: Histogram of politics data OTC score

4.4 BIG DATA VERACITY: DGS MODEL

In this section, we present the results of evaluation of the quantitative measures: Diffusion, Geographic, Spam Indices (DGS) applied to the flu, food poisoning and politics data. As mentioned previously, the DGS model is defined by the following formulae (refer to section 2.1.4):

$$\text{Diffusion Index} = \frac{\#\,\text{Unique Users}}{\text{Total tweets}},$$

$$\text{Geographic Spread Index} = \frac{\#\,\text{Unique Location}}{\text{Total tweets}} \text{ and}$$

$$\text{Spam Index} = \frac{\sum_{\text{over unique users}} \frac{1}{\text{unique user tweet count}}}{\text{Total tweets}}$$

Tables 7, 9, and 11 show the computed DGS model scores for the flu topics (refer to section 4.1.2), Food poisoning topics (refer to section 4.1.2) and the Politics Data topics (refer to section 4.1.2) respectively. Figures 9, 10, and 11 show their 3D plots.

From Figure 9, 10 and 11, we have found out the distance of each point from (1, 1, 1). If the distance from (1, 1, 1) is less, then there will be a higher degree of certainty and a higher degree of veracity. If the distance from (1, 1, 1) is higher, it will show a higher degree of uncertainty and thus lower degree of veracity. Based on this we can estimate the degree and ordering of the veracities of the topics. Tables 8, 10 and 12 show the computed distances of each point from (1, 1, 1) in 3D plots.

Table 7: Flu data DGS model

Topic	Diffusion Index	Geographic Index	Spam Index
1	0.32345	0.07823	1.0063e-13
2	0.08028	0.02046	2.8402e-13
3	0.19226	0.05098	1.1633e-13

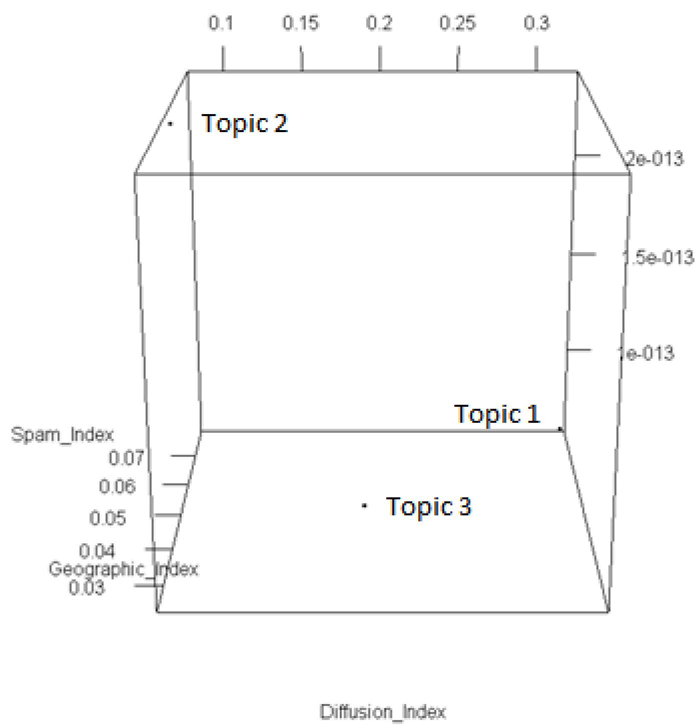


Figure 9: 3D plot of flu data DGS model

Table 8: Flu data DGS model distances

Topic	Distance of each Topic from point (1,1,1)
Topic 1	1.62
Topic 2	1.72
Topic 3	1.69

Table 9: Food Poisoning Data DGS model

Topic	**Diffusion Index**	**Geographic Index**	**Spam Index**
1	0.3536	0.01	1.79e-10
2	0.2272	0.0067	1.69e-11
3	0.4659	0.013	5.85e-11

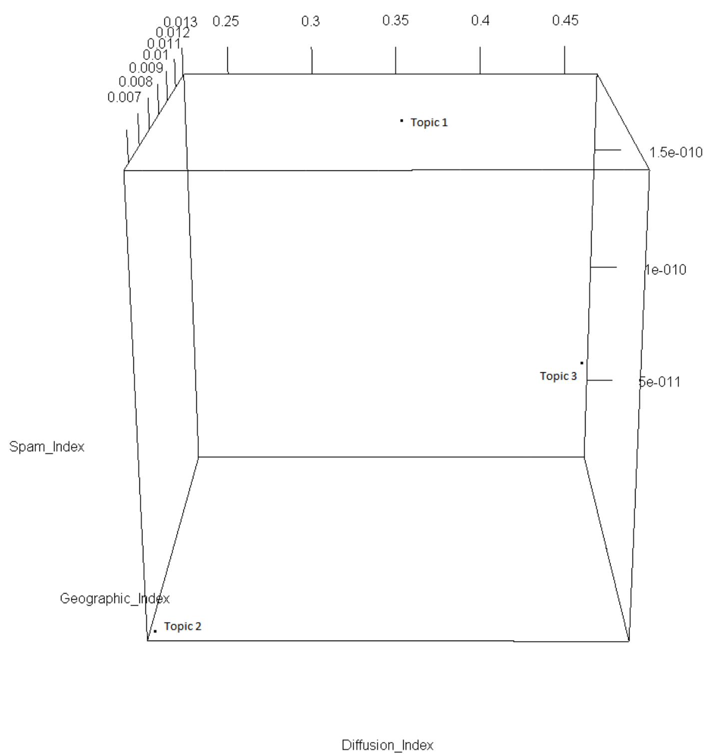


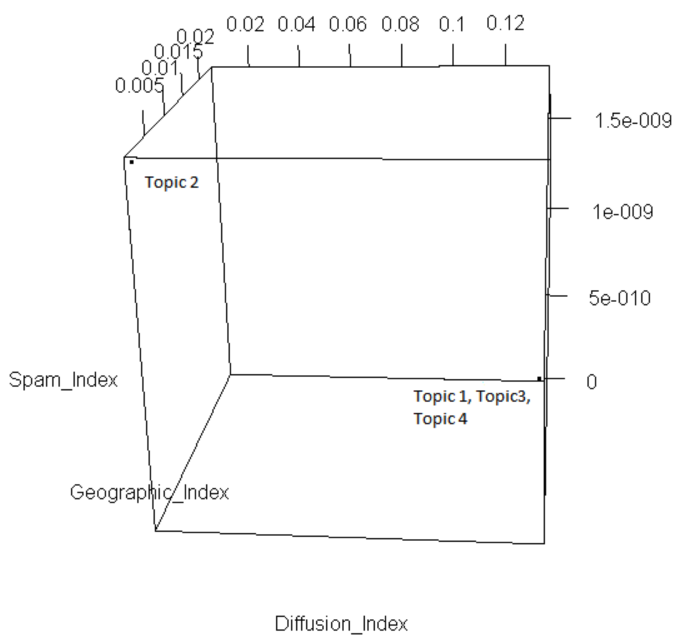
Figure 10: 3D plot of Food Poisoning data DGS model

Table 10: Food Poisoning data DGS model distances

Topic	Distance of each Topic from point (1,1,1)
Topic 1	1.61
Topic 2	1.68
Topic 3	1.51

Table 11: Politics Data DGS model

Topic	**Diffusion Index**	**Geographic Index**	**Spam Index**
1	0.1348	0.0237	8.31e-12
2	0.007021	0.0006385	1.75e-09
3	0.1348	0.0237	8.17e-12
4	0.1348	0.0237	8.80e-12



Politics Data 3-D Graph for Diffusion, Geographic, Spam Index based on topics.

Figure 11: 3D plot of Politics data DGS model

Table 12: Politics data DGS model distances

Topic	Distance of each Topic from point (1,1,1)
Topic 1	1.7133
Topic 2	1.732
Topic 3	1.7133
Topic 4	1.7133

4.5 VERACITY MEASURES COMPARISON BASED ON TOPIC RANKING

Based on analysis done on Veracity measures in previous sections, following is the topic analysis for all 3 datasets using topic rankings for comparison. The model values are computed based on the set of all tweets during the analysis period. All the ranking of topics for the flu, food poisoning, and politics data is done in increasing order of the model values.

Flu data results:

The ranking of the topics is Topic 1, Topic 2 and Topic 3 for Entropy measure and also for the OTC model (refer to section 4.2, Table 3, Figure 3 and Table 6, Figure 6). While, the ranking of topics is Topic 2, Topic 3 and Topic 1 for DGS model (refer to section 4.4, Table 8). This shows that OTC model and Entropy measure almost match. At the same time, they differ in ranking from the DGS model.

Food Poisoning data results:

For the OTC model, the ranking of topics is Topic 2, Topic 1 and Topic 3 (refer to Section 4.3, Table 6, and Figure 7). The ranking in case of Entropy measure is Topic 3, Topic 2 and Topic 1 (refer to section 4.2, Table 4, Figure 4). The ranking of topics for DGS model is Topic 3, Topic 1 and Topic 2 (refer to section 4.4, Table 10). The results show that the three models do not agree on the ranking of topics.

Politics data results:

The ranking of topics for OTC model is Topic 3, Topic 1, Topic 4 and Topic 2 (refer to section 4.3, Table 6, Figure 8). The ranking in case of Entropy measure is Topic 1, Topic 4, Topic 3 and Topic 2 (refer to section 4.2. Table 4, Figure 5). While the ranking of topics for the DGS model is Topic 2, and with other topics, Topic 1, Topic 3 and Topic 4 having the same ranking (refer to section 4.4, Table 12). The results show that the three models do not agree on the ranking of topics. The conclusion is that the three models reflect different properties of the tweets and so cannot be used as corroborating evidence for the pair of models.

The different models do not agree on the ordering of the topics in all the data domains considered. However, their relative values seem to agree on the level of veracity inherent in the data and the placement in the veracity spectrum. To gain further insight, we also conduct a different analysis by representing data as time series.

The next section deals with Time Series analysis of veracity measures done on the three datasets.

4.6 TIME SERIES ANALYSIS OF VERACITY MEASURES

Time Series analysis is performed in order to study different statistics and trends of data for different models and their correlation for each topic. The process deals with calculating the scores of Entropy, OTC model and DGS model for each day and plotting the graph of each topic for all three models.

4.6.1 FOOD POISONING DATA: TIME SERIES ANALYSIS

Figures 12, 13 and 14 are the time series graph of Food Poisoning Data for Topic 1, Topic 2 and Topic 3 respectively (refer to section 4.1.2).

The relationship between the models is calculated by performing analysis of variance (ANOVA) and by calculating the correlation coefficient. ANOVA deals with performing statistical hypothesis testing on sample data and testing the results from the null hypothesis. The test results are statistically significant if it is unlikely to have occurred by chance, that is, if the probability (p-value) is less than the significance level and the F-value is greater than F-critical then it leads to rejection of the null hypothesis. The null hypothesis considered in this case, means all three models (groups) are the same. Then the alternate hypothesis would be at least one of the mean is different from the mean of another model. The significance level considered is 0.05.

Table 13: Food Poisoning topics p-value and F value

Topic Id	p-value	F value
Topic 1	1.1E-148	5691.78
Topic 2	6.9E-144	5261.38
Topic 3	2.9E-149	5792.25

In Table 13, all the p-values are less than 0.05 (significance level) and F-values are greater than 3.05 (F-critical) and so we reject the null hypothesis. So, there is a possibility that at least one of the mean of a particular model is different from the means of other models.

To find the difference between the 3 models, the correlation coefficient is calculated for each topic. Table 14 represent the correlation coefficient scores and the relationship between the models for Topic 1, Topic 2 and Topic 3 respectively for food poisoning data.

If the correlation coefficient (r) lies between ±0.5 and ± 1, then it is said to be a strong correlation, if the r value lies between ±0.3 and ±0.49, then it is said to be a medium correlation and if the r value lies below ±0.29, then it is a weak correlation. There is no correlation if the value is zero.

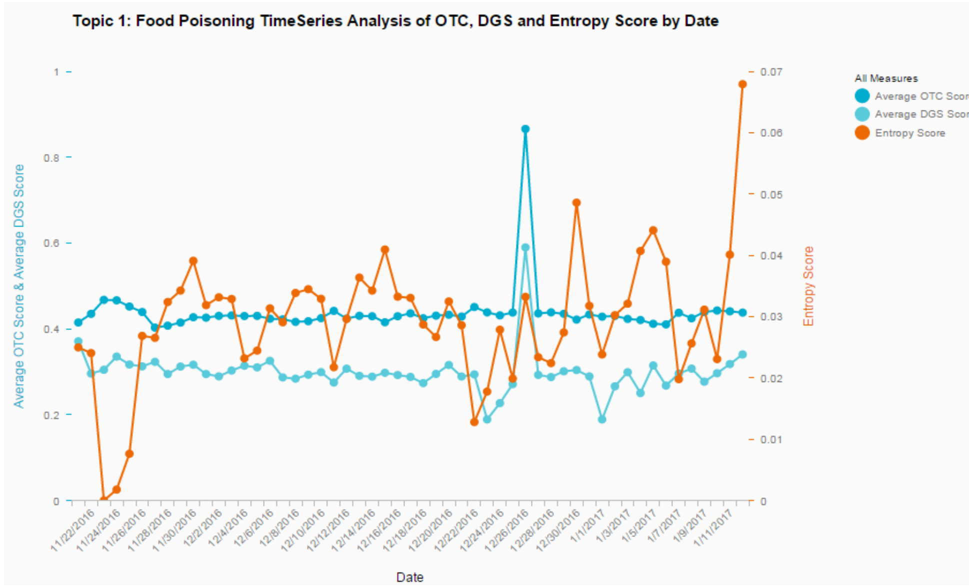


Figure 12: Time series graph of Topic 1 Food Poisoning data

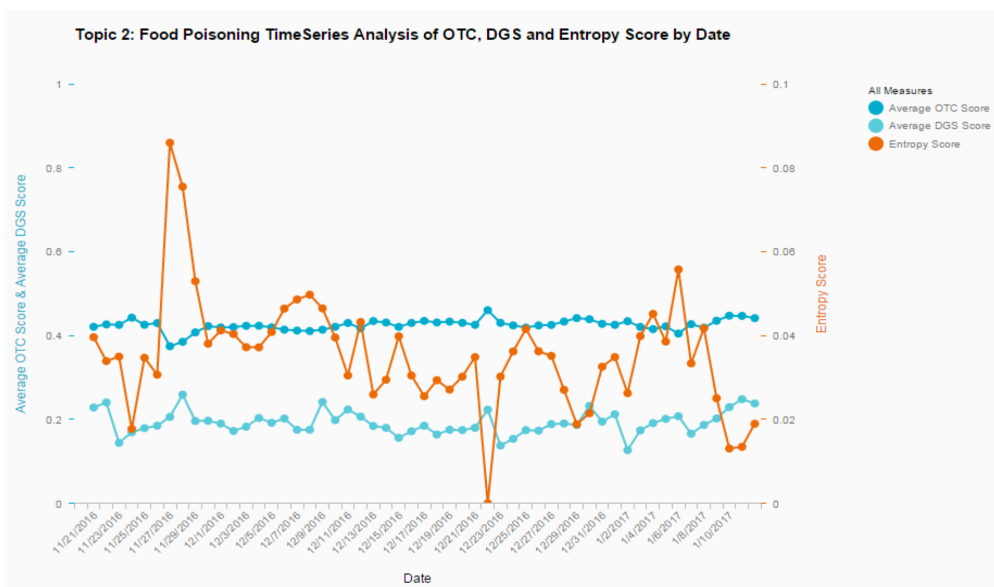


Figure 13: Time series graph of Topic 2 Food Poisoning data

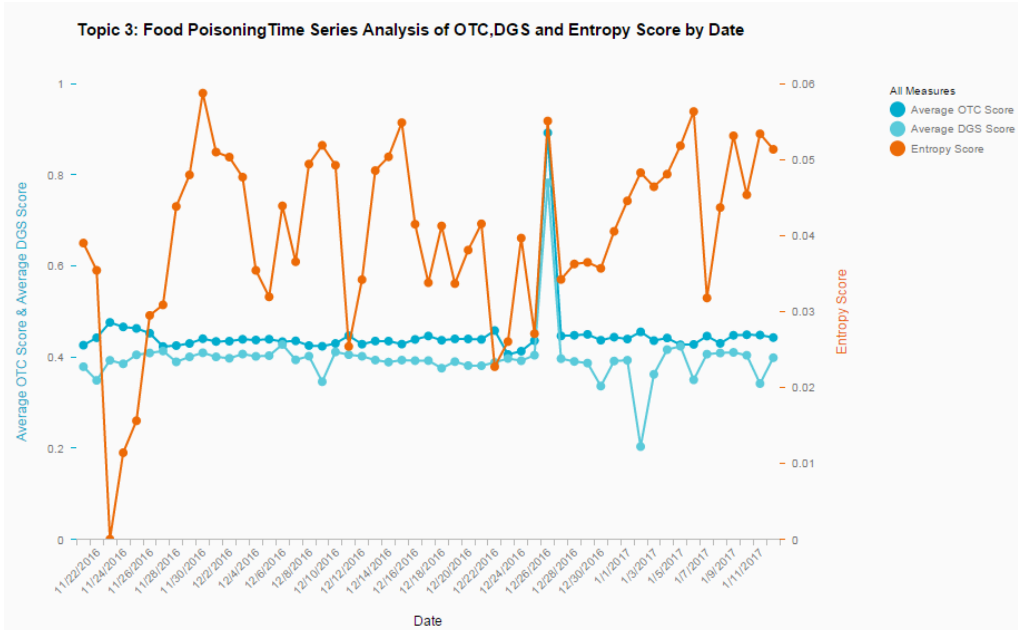


Figure 14: Time series graph of Topic 3 Food Poisoning data

Table 14: Food Poisoning Data Correlation coefficient for topics

Topic	Pairing of 2 models	Correlation coefficient (r)	Relationship
1	OTC and DGS	0.002	Weak positive correlation
2	OTC and DGS	-0.057	Weak negative correlation
3	OTC and DGS	-0.146	Weak negative correlation
1	Entropy and DGS	0.103	Weak positive correlation
2	Entropy and DGS	0.057	Weak positive correlation
3	Entropy and DGS	-0.122	Weak negative correlation
1	Entropy and OTC	-0.5675154	Strong negative correlation
2	Entropy and OTC	-1	Strong negative correlation
3	Entropy and OTC	-0.49361	Strong negative correlation

4.6.2 POLITICS DATA: TIME SERIES ANALYSIS

Figure 15, 16, 17 and 18 are the time series graph of Politics Data for Topic 1, Topic 2, Topic 3 and Topic 4 respectively. (refer to section 4.1.2).

The difference in the models is computed through ANOVA and through finding the correlation coefficient similar to section 4.6.1

Table 15: Politics topics p-value and F value

Topic Id	p-value	F value
Topic 1	2.117E-36	1709.08
Topic 2	2.99E-09	35.55
Topic 3	2.19E-30	353.23
Topic 4	4.59E-31	442.23

In Table 15, all the p-values are less than 0.05(significance level) and F-values are greater than 3.25(F-critical) and so we reject the null hypothesis. So, there is a possibility that at least one of the mean of a particular model is different from the means of other models.

Table 16 represent the correlation coefficient scores and the relationship between the models for Topic 1, Topic 2, Topic 3 and Topic 4 respectively for politics data.

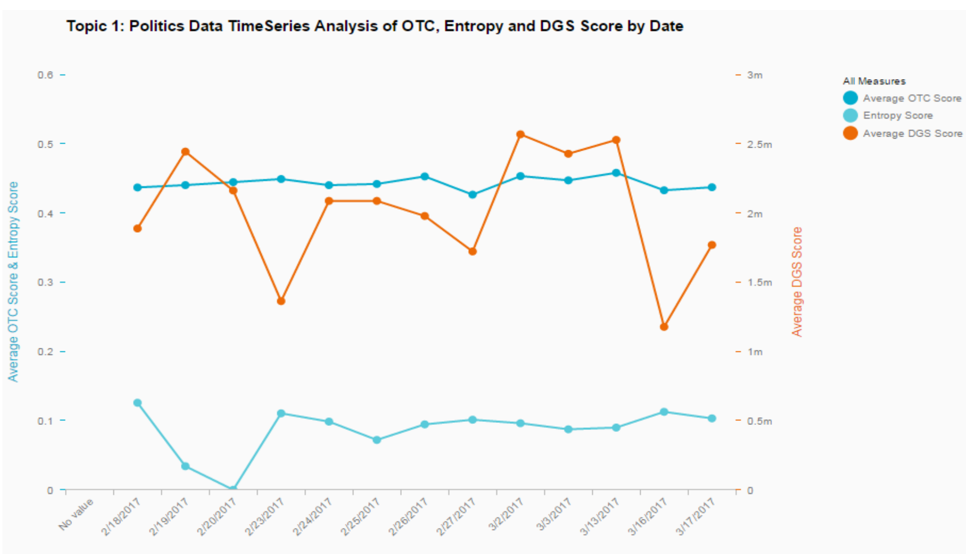


Figure 15: Time series graph of Topic 1 Politics data

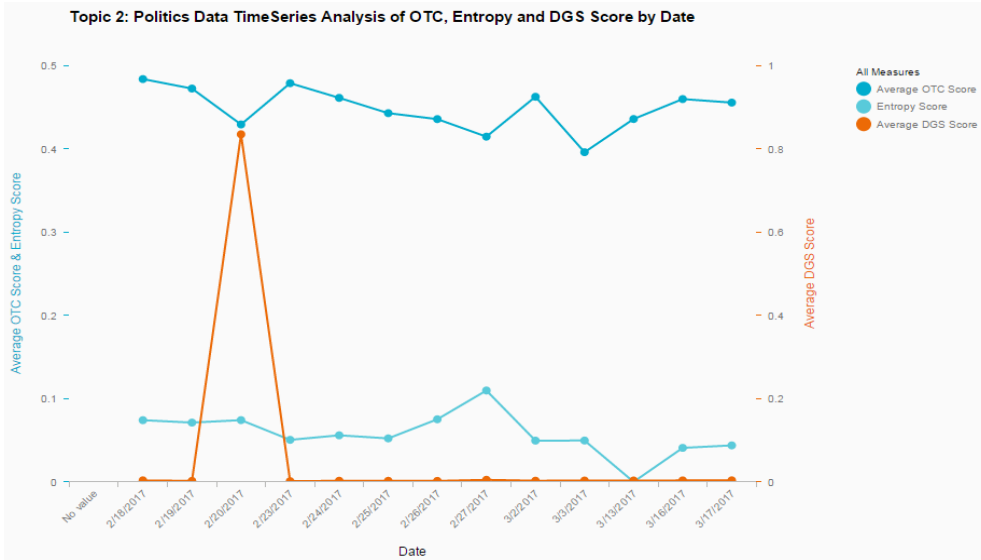


Figure 16: Time series graph of Topic 2 Politics data

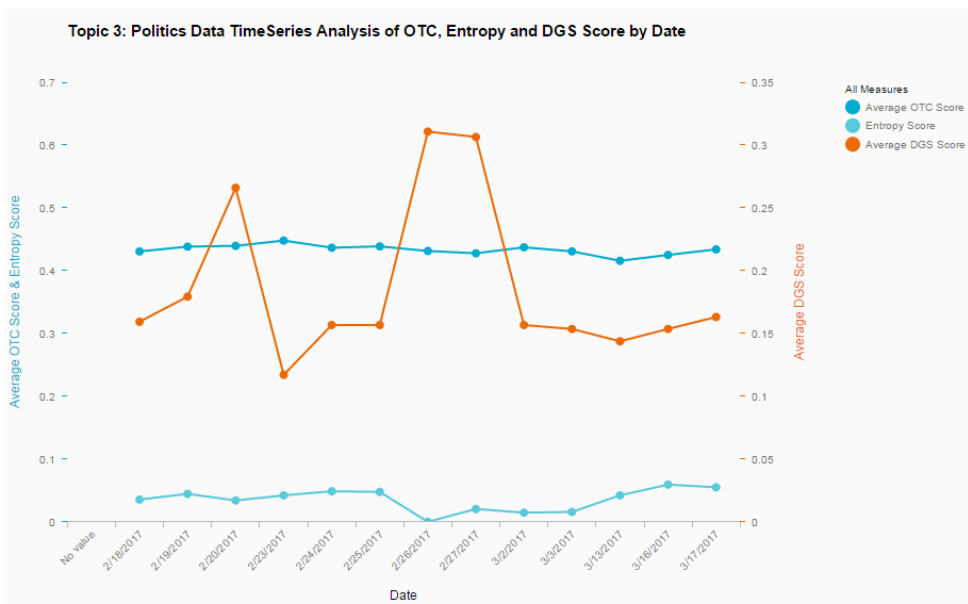


Figure 17: Time series graph of Topic 3 Politics data

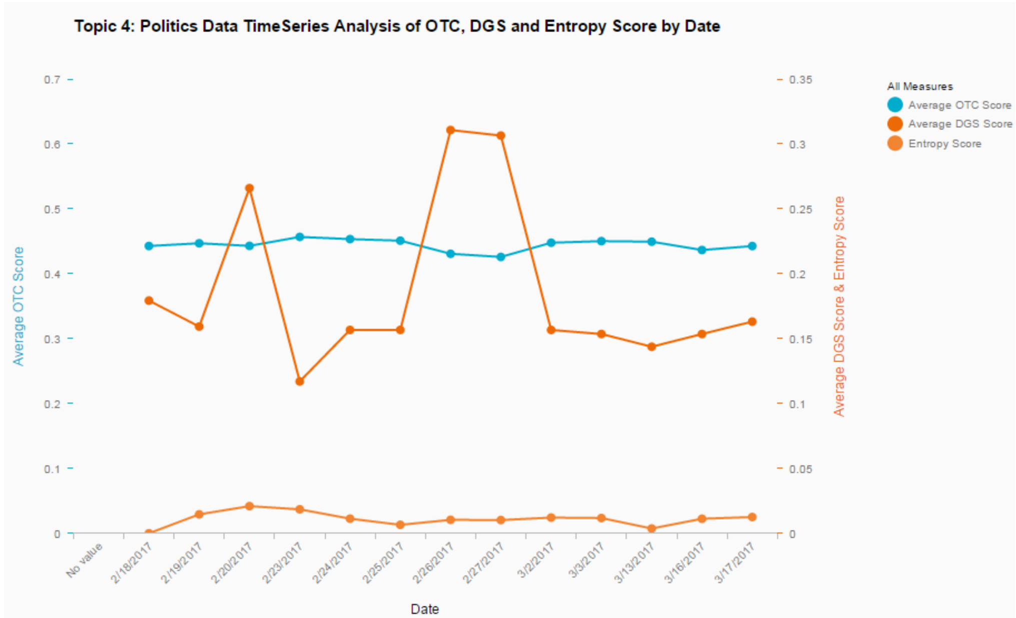


Figure 18: Time series graph of Topic 4 Politics data

Table 16: Politics Data Correlation coefficient for topics

Topic	Pairing of 2 models	Correlation coefficient (r)	Relationship
1	OTC and DGS	0.545	Strong positive correlation
2	OTC and DGS	-0.224	Weak negative correlation
3	OTC and DGS	-0.11	Weak negative correlation
4	OTC and DGS	-0.097	Weak negative correlation
1	Entropy and DGS	-0.445	Weak negative correlation
2	Entropy and DGS	0.198	Weak positive correlation
3	Entropy and DGS	-0.57	Weak negative correlation
4	Entropy and DGS	0.097	Weak positive correlation
1	Entropy and OTC	-0.1308	Weak negative correlation
2	Entropy and OTC	-0.109	Weak negative correlation
3	Entropy and OTC	0.037	Weak positive correlation
4	Entropy and OTC	0.116	Weak positive correlation

4.6.3 FLU DATA: TIME SERIES ANALYSIS

Figure 19, 20 and 21 are the time series graph of Flu Data for Topic 1, Topic 2 and Topic 3 respectively. (refer to section 4.1.2).

The difference in the models is computed through ANOVA and through finding the correlation coefficient similar to section 4.6.1

Table 17: Flu topics p-value and F value

Topic Id	p-value	F value
Topic 1	7.8E-114	805.61
Topic 2	3.9E-171	2467.50
Topic 3	7.9E-207	4582.13

In Table 17, all the p-values are less than 0.05(significance level) and F-values are greater than 3.03(F-critical) and so we reject the null hypothesis. So, there is a possibility that at least one of the mean of a particular model is different from the means of other models.

Table 18 represent the correlation coefficient scores and the relationship between the models for Topic 1, Topic 2 and Topic 3 respectively for politics data.

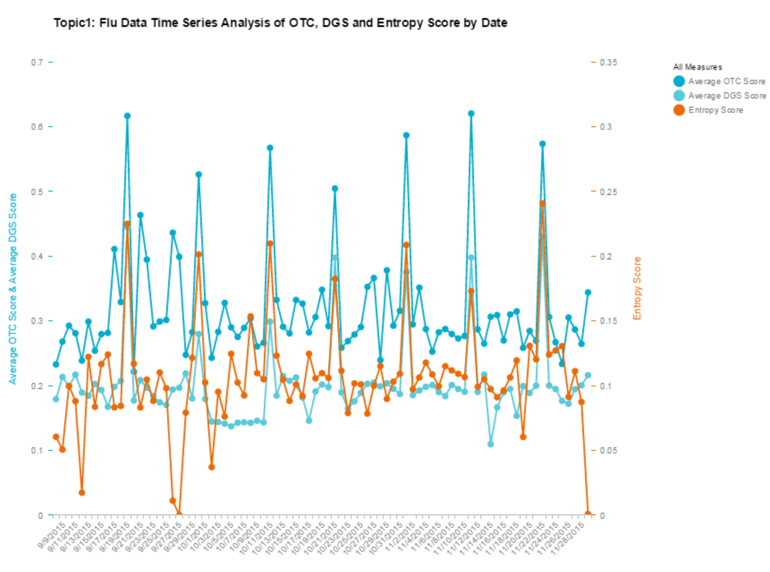


Figure 19: Time series graph of Topic 1 Flu data

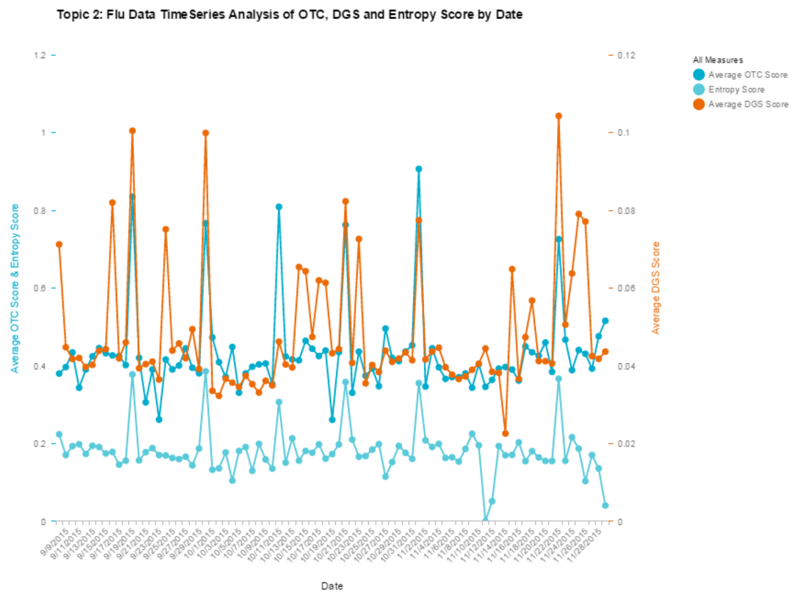


Figure 20: Time series graph of Topic 2 Flu data

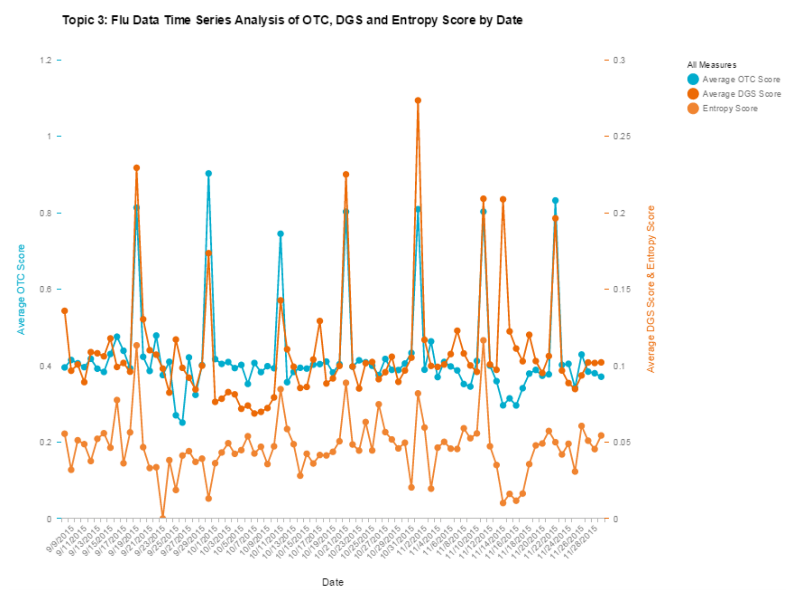


Figure 21: Time series graph of Topic 3 Flu data

Table 18: Flu Data Correlation coefficient for topics

Topic	Pairing of 2 models	Correlation coefficient (r)	Relationship
1	OTC and DGS	0.157	Weak positive correlation
2	OTC and DGS	NA	no correlation

3	OTC and DGS	NA	no correlation
1	Entropy and DGS	-0.125	Weak negative correlation
2	Entropy and DGS	NA	no correlation
3	Entropy and DGS	NA	no correlation
1	Entropy and OTC	-0.211	Weak negative correlation
2	Entropy and OTC	-0.175	Weak negative correlation
3	Entropy and OTC	0.136	Weak positive correlation

4.7 INFERENCE

Flu data results:

Flu data results through topic ranking show that OTC model and Entropy measure almost match with the difference in ranking of Topics for DGS model.

From Section 4.6, Figures 16, 17 and 18 show the daily variations of the three measures in one graph for each topic of the flu dataset. Table 18 shows the correlations and the implied relationships. While the daily graphs demonstrate the agreement of the models on some days, the correlations show more disagreement than agreement. The intermodal comparisons do not provide any corroborating information. Repeating the time series analysis for all the other topics in the politics dataset reveals the same outcomes.

Moreover, according to CDC Data statistics for June 2015 - November 2015 there was 206661 Influenza-related illness visits in the hospital which is 1.24% of a total number of visits related to any kind of illness in hospitals. Of this 1.24% of Influenza-related illness visits, 1.05%

that is 2169 number of cases is Influenza positive tests reported by CDC. The maximum number of Influenza positive tests are found in the month of January (2015) and the number keeps on decreasing with October (2015) month having least number of cases. (Refer to Table 19, row 25). This shows that the tweets used for analysis fall into a non-flu season. So the scores are consistent and do not show a flu epidemic.

Food Poisoning data results:

Food Poisoning data results through topic ranking show that OTC model and DGS model match with the difference in ranking of topics for Entropy measure for food poisoning data.

From Section 4.6, Figures 12, 13 and 14 show the daily variations of the three measures in one graph for each topic of the food poisoning dataset. Table 14 shows the correlations and the implied relationships. While the daily graphs demonstrate the agreement of the models on some days, the correlations show more disagreement than agreement. As in the previous case, the intermodal comparisons do not provide any corroborating information.

Politics data results:

Politics data results through topic ranking show that OTC model, Entropy measure and DGS model.

From Section 4.6, Figures 15, 16 and 17 show the daily variations of the three measures in one graph for each topic of the politics dataset. Table 16 shows the correlations and the implied relationships. Daily graphs show the agreement of models on some days, the correlations here too show more disagreement with only OTC and DGS model show positive correlation for topic 1.

Finally, looking at the results of all three datasets, it can be concluded that OTC model, Entropy measure and DGS model do not show the strong correlation among themselves. The reason that the measures are not unanimous most likely is because of the way the measures are computed. OTC model performs sentiment analysis but needs external information for computation, Entropy is computed by performing sentiment analysis on tweets text without the need for external information. It estimates veracity based on the bag of words and topic model as the basis. DGS model is computed from the tweet themselves using users count and geographic location of the tweets without using external resources.

CHAPTER V

CONCLUSION

Micro-blogging sites like Twitter have become the source of information where people post their real-time experiences and their opinions on various day-to-day issues which can be used to predict and analyze the data. This information sometimes leads to spread of untrue information and has an influence on society. In this thesis, we have proposed Entropy as a measure of veracity and compared its reliability against the previously published measures. The measures are evaluated on the basis of different topics of the data. The topics are defined by words taken from government website based on various contexts of a particular data domain. After comparing the uncertainty of the tweets, our analysis shows some evidence that the model values are dependent on the approaches on which they are based. OTC measure is calculating the sentiments in text along with external sources, entropy measure is calculated through tweet text without any external sources. DGS measure deals with evaluating the accuracy of data from tweet text, users and geographic location without any external sources.

The computed model results do not agree on the topic rankings. However, they place the topics in a veracity spectrum in a consistent manner. Flu data available from CDC for the time period corresponding to Twitter flu data period are used for model validation. We interpret the tweets as indicator of flu. The data period analyzed is not considered flu season and the computed model values are not high. So, the data seems to validate the topic placement in the veracity spectrum by the models. No official data was available corresponding to the timeframe of the other datasets.

Verifying the other datasets with external data is proposed as future work. Also, as entropy computes the veracity of topics based on topic model (bag-of-words), other topic modeling algorithms can be explored as future work.

Table 19: External Links

Sr No	Links
(1)	Apache Hadoop, http://hadoop.apache.org/
(2)	Blob classes, http://textblob.readthedocs.io/en/dev/api_reference.html
(3)	CDC, http://www.cdc.gov/flu/about/disease/symptoms.htm[09-06- 2015], http://www.cdc.gov/flu/about/qa/hospital.htm, http://www.cdc.gov/flu/about/disease/us_flu-related_deaths.htm
(4)	CDC, http://www.cdc.gov/flu/pdf/freeresources/updated/treating_flu.pdf [09-06- 2015]
(5)	Downloading of TextBlob 0.11.1 file, https://pypi.python.org/pypi/textblob
(6)	Flu.gov, http://www.flu.gov/symptoms-treatment/symptoms/ [09-06- 2015]
(7)	Flu.gov, http://www.flu.gov/symptoms-treatment/treatment/ [09-06- 2015]
(8)	Installation of TextBlob, http://textblob.readthedocs.io/en/dev/install.html
(9)	Mayo Clinic, http://www.mayoclinic.org/diseases-conditions/flu/basics/symptoms/con-20035101 [09-06-2015]
(10)	Merriam-Webster Online Dictionary. (2009). Retrieved 10 February 2009, www.merriam-webster.com
(11)	Moffit, J.S.Giboney, 2012 http://splice.cmi.arizona.edu/
(12)	Pattern Module, http://www.clips.ua.ac.be/pattern
(13)	Sentiment Lexicon, https://github.com/sloria/TextBlob/blob/dev/textblob/en/en-sentiment.xml
(14)	TextBlob: Simplified Text Processing, https://textblob.readthedocs.io/en/latest/#features
(15)	TextBlob Sentiment: Calculating Polarity and Subjectivity, http://planspace.org/20150607-textblob_sentiment/
(16)	TextBlob Documentation, https://media.readthedocs.org/pdf/textblob/dev/textblob.pdf
(17)	TextBlob Sentiment Analyzers, http://textblob.readthedocs.io/en/dev/api_reference.html#textblob.en.sentiments.PatternAnalyzer
(18)	WebMD, http://www.webmd.com/cold-and-flu/flu-guide/adult-flu-symptoms
(19)	WebMD, http://www.webmd.com/cold-and-flu/flu-guide/is-it-cold-flu [09-06- 2015]
(20)	WebMD, http://www.webmd.com/cold-and-flu/flu-guide/flu-treatment [09-06- 2014]
(21)	HDFS Architecture Guide https://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf
(22)	Apache Flume https://flume.apache.org/
(23)	CDC, https://www.cdc.gov/foodsafety/foodborne-germs.html
(24)	CDC, https://www.cdc.gov/foodsafety/diseases/index.html
(25)	CDC, https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

REFERENCES

[1] Seth A and Sharma, Aneesh and Gupta, Pankaj and Lin, Jimmy Myers, *Information Network or Social Network? The structure of Twitter Follow Graph.*, 2014.

[2] Anabel and Martin, Kim and McCay-Peet, Lori Quan-Haase, "Networks of digital humanities scholars: The informational and social uses and gratifications of Twitter," *Big Data \& Society*, vol. 2, p. 2053951715589417, 2015.

[3] Mylynn Felt, "Social media and the social sciences: How researchers employ Big Data analytics," *Big Data & Society*, vol. 3, p. 2053951716645828, 2016.

[4] John and Reinsel, David Gantz, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the future*, vol. 2007, pp. 1--16, 2012.

[5] Yuri and Ngo, Canh and de Laat, Cees and Membrey, Peter and Gordijenko, Daniil Demchenko, "Big security for big data: addressing security challenges for the big data infrastructure," , Trento, Italy, 2013, pp. 76--94, Workshop on Secure Data Management, Springer.

[6] IBM Big Data\&Analytics Hub, "The Four V's of Big Data," *2015]. http://www. ibmbigdatahub. com/infographic/four-vsbig-data*, 2013.

[7] Normala Binti Che and Ishak, Iskandar Bin and Sidi, Fatimah and Affendey, Lilly Suriani and Mamat, Ali and others Eembi, "A Systematic Review on the Profiling of Digital News Portal for Big Data Veracity," *Procedia Computer Science*, vol. 72, pp. 390--397, 2015.

[8] Shen and Kaynak, Okyay Yin, "Big Data for Modern Industry: Challenges and Trends [Point of View]," *Proceedings of the IEEE*, vol. 103, pp. 143--146, 2015.

[9] B Marr, "Why only one of the 5 Vs of big data really matters," *IBM Big Data \& Analytics Hub. Available online at www. ibmbigdatahub. com/blog/whyonly-one-5-vs-big-data-really-matters (last accessed February 29, 2016*, 2015.

[10] Tatiana and Rubin, Victoria L Lukoianova, "Veracity roadmap: Is big data objective, truthful and credible?," *Advances in Classification Research Online*, vol. 24, pp. 4--15, 2014.

[11] Kumar TK and Kammarpally, Prashanth and George, KM Ashwin, "Veracity of information in twitter data: A case study," in *2016 International Conference on Big Data and Smart Computing (BigComp)*, 2016, pp. 129--136.

[12] Johannes and Richthammer, Christian and Hassan, Sabri and Pernul, Gunther Sanger, "Trust and big data: a roadmap for research," in *2014 25th International Workshop on Database and Expert Systems Applications*, 2014, pp. 278--282.

[13] Reed, Colorado, "Latent Dirichlet Allocation: Towards a Deeper Understanding," 2012.

[14] David M and Ng, Andrew Y and Jordan, Michael I Blei, "Latent Dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993--1022, 2003.

[15] Thomas K and Foltz, Peter W and Laham, Darrell Landauer, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, pp. 259--284, 1998.

[16] Seungjin Choi, "Probabilistic Latent Semantic Analysis," 2011.

[17] Thomas Hofmann, "Probabilistic latent semantic analysis," , 1999, pp. 289--296.

[18] Wei and Liu, Xin and Gong, Yihong Xu, "Document clustering based on non-negative matrix factorization," , 2003.

[19] David M and Lafferty, John D Blei, "A correlated topic model of science," *The Annals of Applied Statistics*, pp. 17--35, 2007.

[20] Claude Elwood Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, pp. 3--55, 2001.

[21] Ralph VL Hartley, "Transmission of information1," *Bell System technical journal*, vol. 7, pp. 535--563, 1928.

[22] Denis McQuail, *McQuail's mass communication theory*.: Sage publications, 2010.

[23] Paul K Moser, *Philosophy after objectivity: Making sense in perspective*. New York: Oxford University Press on Demand, 1993.

[24] Richard Rorty, *Objectivity, relativism, and truth: philosophical papers*.: Cambridge University Press, 1991, vol. 1.

[25] Graeme Hirst, *Views of Text Meaning in Computational Linguistics: Past, Present, and Future*., 2007.

[26] David B and Burgoon, Judee K Buller, "Interpersonal deception theory," *Communication theory*, vol. 6, pp. 203--242, 1996.

[27] Lina and Burgoon, Judee K and Nunamaker, Jay F and Twitchell, Doug Zhou, "Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communications," *Group decision and negotiation*, vol. 13, pp. 81--106, 2004.

[28] Subbalakshmi K. Chandramouli R., "Text Analytics: Deception Detection and Gender Identification from Text," *Digital Investigation: The International Journal of Digital Forensics & Incident Response*, vol. 8(1), July 2011.

[29] Myle and Choi, Yejin and Cardie, Claire and Hancock, Jeffrey T Ott, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2011, pp. 309--319, Association for Computational Linguistics.

[30] Victoria L and Vashchilko, Tatiana Rubin, "Extending information quality assessment methodology: A new veracity/deception dimension and its measures,"

Proceedings of the American Society for Information Science and Technology, vol. 49, pp. 1--6, 2012.

[31] BJ and Tseng, Hsiang Fogg, "The elements of computer credibility," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*., 1999, pp. 80--87.

[32] Mark Davies, "The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. 2008," *Available online a t http://www. americancorpus. org*, February 2009.

[33] Judee K and Blair, JP and Qin, Tiantian and Nunamaker Jr, Jay F Burgoon, "Detecting deception through linguistic analysis," in *International Conference on Intelligence and Security Informatics*, 2003, pp. 91--101.

[34] Jeffrey T and Curry, Lauren E and Goorha, Saurabh and Woodworth, Michael Hancock, "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication," *Discourse Processes*, vol. 45, pp. 1--23, 2007.

[35] Matthew L and Pennebaker, James W and Berry, Diane S and Richards, Jane M Newman, "Lying words: Predicting deception from linguistic styles," *Personality and social psychology bulletin*, vol. 29, pp. 665--675, 2003.

[36] Leticia C and Rosso, Paolo Cagnina, "Classification of deceptive opinions using a low dimensionality representation," in *6TH WORKSHOP ON COMPUTATIONAL APPROACHES TO SUBJECTIVITY, SENTIMENT AND SOCIAL MEDIA ANALYSIS WASSA 2015*, Lisboa, Portugal, 2015, p. 58.

[37] Tom De and Daelemans, Walter Smedt, "Pattern for Python," *Journal of Machine Learning Research*, vol. 13, pp. 2063--2067, 2012.

[38] Tom De Smedt, *Modeling Creativity: Case Studies in Python.*: University Press Antwerp, 2013.

[39] Valentin and Hofmann, Katja Jijkoun, "Generating a non-English subjectivity lexicon: relations that matter," , 2009, pp. 398--405.

[40] Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry Page, "The PageRank citation ranking: bringing order to the web.," 1999.

[41] Tim Van de Cruys, "Mining for Meaning: The Extraction of Lexicosemantic Knowledge from Text," *Groningen Dissertations in Linguistics*, vol. 82, 2010.

[42] Grigori and Gelbukh, Alexander and G, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Computaci*

[43] Bo and Lee, Lillian Pang, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 2004, p. 271, Association for Computational Linguistics.

[44] Andrea H and Moore, Kathleen A and Johnson, Nichloas J Tapia, "Beyond the trustworthy tweet: A deeper understanding of microblogged data use by disaster response and humanitarian relief organizations," , 2013, pp. 770--778.

[45] Jeremy and Gadepally, Vijay and Michaleas, Pete and Schear Nabil and Varia, Mayank and Yerukhimovich, Arkady and Cunningham, Robert K Kepner, "Computing on masked data: a high performance method for improving big data veracity," in *High Performance Extreme Computing Conference (HPEC), 2014 IEEE*, 2014, pp. 1--6.

[46] Saima and Chelmis, Charalampos and Prasanna, Viktor Aman, "Addressing data veracity in big data applications," in *Big Data (Big Data), 2014 IEEE International Conference on*, 2014, pp. 1--3.

[47] Xinzhi and Luo, Xiangfeng and Liu, Huiming Wang, "Measuring the veracity of web event via uncertainty," *Journal of Systems and Software*, vol. 102, pp. 226--236, 2015.

[48] Xizhao Wang and Yulin He, "Learning from uncertainty for Big Data," vol. 2, no. 2, pp. 26-31, 2016.

[49] Apoorv and Xie, Boyi and Vovsha, Ilia and Rambow, Owen and Passonneau, Rebecca Agarwal, "Sentiment analysis of twitter data," in *Association for Computational Linguistics*, 2011, pp. 30--38.

APPENDICES

1. Twitter Data Streaming Configuration file

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sources.Twitter.consumerKey = B5cZu4txtGCrdkq8ShCWoVcju

TwitterAgent.sources.Twitter.consumerSecret = 339n0j3g2HMF0qV9zAMKp95XyThLPOjSQhfiK6nv5MIZ3sqeAR

TwitterAgent.sources.Twitter.accessToken = 3379110614-BNtMtUKNKNECFUdfh3WDugZ0UflHU03gaBD1Pna

TwitterAgent.sources.Twitter.accessTokenSecret = Rzv616mODXltA4Y145rRpQ2El49yZCixnj39gYIbbaYUe

TwitterAgent.sources.Twitter.keywords = diarrhea,fever,abdominal pain,abdominal cramps,vomiting,bloody diarrhea,muscle ache,nausea,headache,stiff neck,confusion,convulsion,chills,watery diarrhea,stomach cramps,weight loss,slight fever,greasy stools,gas and bloating,double vision,blurred vision,drooping eyelids,slurred speech,difficulty swallowing,dry mouth,muscle weakness,jaundice,dark-colored urine,light-color stool

```
TwitterAgent.sinks.HDFS.channel = MemChannel

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path = hdfs://hadoop1:9000/paryani/FoodPoisioningData/

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream

TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text

TwitterAgent.sinks.HDFS.hdfs.batchSize = 100

TwitterAgent.sinks.HDFS.hdfs.rollSize = 0

TwitterAgent.sinks.HDFS.hdfs.rollCount = 0

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 10000

TwitterAgent.channels.MemChannel.transactionCapacity = 1000
```

2. Sample JSON Data

{"filter_level":"low","retweeted":false,"in_reply_to_screen_name":null,"possibly_sensiti
ve":false,"truncated":false,"lang":"en","in_reply_to_status_id_str":null,"id":64130298632
9751552,"in_reply_to_user_id_str":null,"timestamp_ms":"1441733521431","in_reply_to
_status_id":null,"created_at":"Tue Sep 08 17:32:01 +0000
2015","favorite_count":0,"place":null,"coordinates":null,"text":"RT @ColleenB123: Also
- diarrhea is
awful.","contributors":null,"retweeted_status":{"filter_level":"low","contributors":null,"t
ext":"Also - diarrhea is
awful.","geo":null,"retweeted":false,"in_reply_to_screen_name":null,"possibly_sensitive
":false,"truncated":false,"lang":"en","entities":{"trends":[],"symbols":[],"urls":[],"hashtag
s":[],"user_mentions":[]},"in_reply_to_status_id_str":null,"id":641141853174128640,"so
urce":"<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for
iPhone<\/a>","in_reply_to_user_id_str":null,"favorited":false,"in_reply_to_status_id":nul
l,"retweet_count":437,"created_at":"Tue Sep 08 06:51:44 +0000
2015","in_reply_to_user_id":null,"favorite_count":2833,"id_str":"641141853174128640"

,"place":null,"user":{"location":"Los
Angeles","default_profile":false,"profile_background_tile":true,"statuses_count":16521,"
lang":"en","profile_link_color":"2FC2EF","profile_banner_url":"https://pbs.twimg.com/p
rofile_banners/267305045/1430708583","id":267305045,"following":null,"protected":fal
se,"favourites_count":24432,"profile_text_color":"333333","verified":true,"description":"
I am Miranda Sings alter ego. I like cookies, documentaries, kittens, and that guy I
married.
http://youtube.com/psychosoprano","contributors_enabled":false,"profile_sidebar_border
_color":"EEEEEE","name":"Colleen
Ballinger","profile_background_color":"1A1B1F","created_at":"Wed Mar 16 17:55:20
+0000
2011","default_profile_image":false,"followers_count":1007677,"profile_image_url_http
s":"https://pbs.twimg.com/profile_images/635317228716564480/KypMTpAG_normal.jp
g","geo_enabled":true,"profile_background_image_url":"http://pbs.twimg.com/profile_ba
ckground_images/602200021/80umi8063weaa53pvk8d.jpeg","profile_background_imag
e_url_https":"https://pbs.twimg.com/profile_background_images/602200021/80umi8063
weaa53pvk8d.jpeg","follow_request_sent":null,"url":null,"utc_offset":-
25200,"time_zone":"Arizona","notifications":null,"profile_use_background_image":true,
"friends_count":8235,"profile_sidebar_fill_color":"EFEFEF","screen_name":"ColleenB1
23","id_str":"267305045","profile_image_url":"http://pbs.twimg.com/profile_images/635
317228716564480/KypMTpAG_normal.jpg","listed_count":1536,"is_translator":false},"
coordinates":null},"geo":null,"entities":{"trends":[],"symbols":[],"urls":[],"hashtags":[],"
user_mentions":[{"id":267305045,"name":"Colleen
Ballinger","indices":[3,15],"screen_name":"ColleenB123","id_str":"267305045"}]},"sour
ce":"<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for
iPhone<\/a>","favorited":false,"in_reply_to_user_id":null,"retweet_count":0,"id_str":"64
1302986329751552","user":{"location":"","default_profile":false,"profile_background_ti
le":false,"statuses_count":2707,"lang":"en","profile_link_color":"3B94D9","profile_bann
er_url":"https://pbs.twimg.com/profile_banners/3130256735/1434982226","id":3130256
735,"following":null,"protected":false,"favourites_count":4441,"profile_text_color":"000
000","verified":false,"description":"Justin Bieber, Ariana Grande, Selena Gomez, and
Colleen Evans mean so much to me. Couldnt ask for better idols! I made this account just
to follow them
:)","contributors_enabled":false,"profile_sidebar_border_color":"000000","name":"Gillia
nEvnsBallinger","profile_background_color":"000000","created_at":"Tue Mar 31
01:06:48 +0000
2015","default_profile_image":false,"followers_count":35,"profile_image_url_https":"htt
ps://pbs.twimg.com/profile_images/627314665668898816/3CvXIrWf_normal.jpg","geo_
enabled":false,"profile_background_image_url":"http://abs.twimg.com/images/themes/th
eme1/bg.png","profile_background_image_url_https":"https://abs.twimg.com/images/the
mes/theme1/bg.png","follow_request_sent":null,"url":"http://evansjourneybegan7-2-
15.com","utc_offset":-25200,"time_zone":"Pacific Time (US &
Canada)","notifications":null,"profile_use_background_image":false,"friends_count":55,
"profile_sidebar_fill_color":"000000","screen_name":"colleenb127","id_str":"313025673
5","profile_image_url":"http://pbs.twimg.com/profile_images/627314665668898816/3Cv
XIrWf_normal.jpg","listed_count":1,"is_translator":false}}

VITA

JYOTSNA PARYANI

COMPUTER SCIENCE

Master of Science

Thesis: A CASE STUDY ON DETERMINING THE BIG DATA VERACITY: A METHOD TO COMPUTE THE RELEVANCE OF TWITTER DATA

Major Field: Computer Science

Biographical:

Education:

Completed the requirements for the Master of Science in Computer Science at Oklahoma State University, Stillwater, Oklahoma in May 2017.

Completed the requirements for the Bachelor of Engineering in Computer Engineering at Pune University, Pune, India in 2012.